

## Data Release Policy for the Cancer Target Discovery and Development (CTD<sup>2</sup>) Network

### Background:

The rapidly evolving sequencing technologies and informatics tools have made it feasible to comprehensively characterize tumor transcriptomes and genomes with reasonable accuracy and cost. The NCI supports the molecular characterization of tumors at the genome level for many cancers as part of the Cancer Genome Characterization Initiative ([CGCI](#)), Therapeutically Applicable Research to Generate Effective Targets ([TARGET](#)) and the Cancer Genome Atlas ([TCGA](#)) (in collaboration with NHGRI) to obtain the detailed information on the repertoire of alterations in a variety of tumors. The next five to ten years will result in a compendium of all possible changes from the projects listed above, as well as the recently initiated International Cancer Genome Consortium. The initial results point very clearly that it are not the individual genes and alterations that are the drivers of cancer, rather the alternations of one or more pathways. The Cancer Target Discovery and Development (CTD<sup>2</sup>) network is a pilot to develop new scientific approaches to accelerate the translation of the genomic discoveries into new treatments. The network emphasizes interaction of laboratories with complementary and unique expertise, including bioinformatics, genome-wide loss of function screening and targeted gain-of-function candidate gene validations, judicious use of mouse-based screens and small molecule high-throughput screens.

### Policy:

CTD<sup>2</sup> is a “community resource project” (ref), which requires rapid data release to enable other research to enhance the Network’s clinical impact. Therefore patenting on the PRIMARY data is discouraged to allow the scientific community easy access and encourage its use.

The data release policy for CTD<sup>2</sup> is consistent with NCI-funded large-scale genomic characterization projects. The data generated by CTD<sup>2</sup> involve a substantial investment of public funds, and are the logical continuation of genomic data sets that will be produced over the next months and years and which are continually expanded by the addition of new tumor types to the aforementioned efforts. To best accomplish the goals of the project, the Institute and society, that is to facilitate research and reduce redundancy by making primary data available to the scientific community in real time, the project members agreed on the following policy:

- *Release of primary (raw) data to the public will occur as soon as data is confirmed and no later than when a manuscript is accepted for publication. The verified data will be posted for use by the CTD<sup>2</sup> network on the website (<http://ctd2.nci.nih.gov>).*
- *The raw data will be accompanied by detailed metadata explaining the conditions in which the experiments were performed, genotypic information on the cell lines or animal models employed, analysis tools and parameters used for data processing, etc.*
- *Data deposited in this portal will be reported in two formats as appropriate:*
  - *If an industry-standard or community-agreed standard exists for the data reported, and the datafile can be analyzed with freely available tools, then that standard will be used.*

*Examples: .CEL files (microarray), .SDF files (cheminformatics), .GCT files (Gene Pattern).*

- *Datasets lacking widely accepted standards (such as RNAi screens), will be shared as data tables (TAB or CSV) which will facilitate their parsing and importing into a wide variety of informatic tools. In those cases, extensive documentation on the file contents and metadata will be also provided to simplify the data file use.*

*Examples: HTS data from UTSW and Broad Institute.*

Several data types will be produced in a variety of biologically relevant experimental models. A summary list of both data elements is provided below. This list is not meant to be comprehensive.

### **1. Biological data types to be shared**

- Small-molecule perturbation data
- RNAi perturbation data
- Networks, interactions, and pathways; inferred from data or predefined from literature
- Genotypic/phenotypic characterization of cell lines used in loss-of-function and gain of function screens:
  - mRNA expression data (MAGE-TAB)
  - Copy number data
  - Mutation data

### **2. Detailed data on experimental objects (metadata)**

- Experiments and analyses (inputs → protocols → outputs)
- Small molecules (names and/or structures)
- Information about phenotypes (e.g., induced by RNAi or small molecules)
- Genes and proteins (including transcript variants)
- Cell lines (including genotypic characterization)
- Description of the mouse models

The CTD<sup>2</sup> data portal (<http://ctd2.nci.nih.gov>) will include a text about the philosophy of the rapid data release policy, “The Responsible use and Publication of Data Generated by the Cancer Target Discovery and Development Network”. The language will be aligned as much as possible to the one used for TCGA and TARGET.

To support the continued prompt public release of large-scale genomic data prior to publication, researchers who plan to prepare manuscripts that would be comparable to the analyses described above, and journal editors who receive such manuscripts, are encouraged to coordinate their independent reports with the project’s publication schedule described above. This may be done by contacting the Project Team.

Researchers are encouraged to use CTD<sup>2</sup> data to publish on the development of novel methods.

NCI does not consider that deposition of data from the CTD<sup>2</sup>, like those from other large-scale genomic projects, into its own or public databases to be the equivalent of publication in a peer-reviewed journal. Therefore, although the data are available to others, the producers still consider

them to be formally unpublished and expect that the data will be used in accord with standard scientific etiquette and practices concerning unpublished data.

The CTD<sup>2</sup> network requests that researchers who use the data generated as part of the project acknowledge it as follows: *“The results published here are in whole or part based upon data generated by Cancer Target Discovery and Development (CTD<sup>2</sup>) network. Information about project, the investigators and institutions that constitute the CTD<sup>2</sup> can be found at <http://ocg.cancer.gov/programs/ctdd.asp> and described in Nature Biotechnology Sep 2010; 28(9):904-906”*. After specific publication of the data by a CTD<sup>2</sup> center(s), the paper should also be referenced.

Meeting presentations of CTD<sup>2</sup> data and analyses by network members are possible and encouraged. We would request that the network members inform the NCI of oral and poster presentations in scientific meetings and fora.

DRAFT